

University of Alberta

A Complete Test of Hulin's Psychometric Theory of Measurement Equivalence on

Translated Tests

by

Shameem Nyla Khaliq



A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment
of the requirements for the degree of Master of Education

Department of Educational Psychology

Edmonton, Alberta

Fall 2000



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-59755-5

University of Alberta

Library Release Form

Name of Author: Shameem Nyla Khaliq

Title of Thesis: A Complete Test of Hulin's Psychometric Theory of Measurement

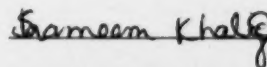
Equivalence on Translated Tests

Degree: Master of Education

Year this Degree Granted: 2000

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



Shameem Nyla Khaliq
4603-113 A Street
Edmonton, Alberta
T6H 1A1

July 19, 2000

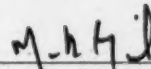
Abstract

Hulin (1987) proposed that the *a* and *b* item parameters might indicate why an item is functioning differentially between examinees on translated tests. He suggested that the *a* parameter indicates differences due to cultural variation and the *b* parameter indicates differences due to translation errors. French Immersion and Francophone students were compared – different culture, same language. French Immersion students, half, randomly selected, wrote in English were compared to the remaining wrote in French – same culture, different language. According to Hulin's theory, there should be more *a* parameter differences when comparing the French Immersion to the Francophone students whereas there should be more *b* parameter differences when comparing the French Immersion students. For the different culture, same language comparison, 8 items displayed differences in the *a* parameter and 3 items displayed differences in the *b* parameter. For the same culture, different language comparison, 7 of the 8 items displayed differences in the *b* parameter and 5 items displayed differences in the *a* parameter. However, two limitations should be noted. First, small samples were used in this study. Second, the results may not be generalizable to monolingual examinees.

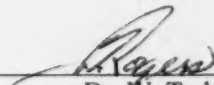
University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled A Complete Test of Hulin's Psychometric Theory of Measurement Equivalence on Translated Tests submitted by Shameem Nyla Khaliq in partial fulfillment of the requirements for the degree of Master of Education.



Dr. Mark J. Gierl



Dr. W. Todd Rogers



Dr. Jeffrey Bisanz

Acknowledgments

I would to acknowledge the support of my committee members Dr. Jeffrey Bisanz, Dr. W. Todd Rogers, and my supervisor Dr. Mark J. Gierl. Jeff, for suggesting that I look into the field of measurement. Todd and Mark, for supporting me throughout the Master of Education program by giving me assistance with the coursework. Mark, for providing me assistance and guidance with this thesis and for making me aware of the possibilities that exist in the area of measurement. I would also like to thank the graduate students of the Centre for Research in Applied Measurement and Evaluation for their constant support and encouragement. Finally, I would like to say thank to my mother, Joan Khaliq, who has always been there for me, no matter what the situation and no matter how stressed I felt.

Table of Contents

I. Introduction	1
II. Methods	10
A. Subjects	10
B. Instrument	12
C. Analyses	13
III. Results	18
A. Different Culture, Same Language	18
B. Same Culture, Different Language	24
IV. Summary and Conclusion	29
A. Limitations	32
B. Improvements	34
C. Implications	34
V. References	39

List of Tables

Table 1. Descriptive Characteristics for French Immersion	
Francophone	44
Table 2. Tests for Model Equivalence Between French Immersion	
and Francophone Students	45
Table 3. Item Parameters for the Mathematics Items - French	
Immersion versus Francophone	46
Table 4. Item Parameters for the Social Studies Items - French	
Immersion versus Francophone	48
Table 5. Descriptive Characteristics for French Immersion	
Students	50
Table 6. Tests for Model Equivalence Between French Immersion	
Students Who Wrote in English and Who Wrote in	
French	51
Table 7. Item Parameters for the Mathematics Items - FIE	
versus FIF	52
Table 8. Item Parameters for the Social Studies Items - FIE	
versus FIF	54

List of Figures

Figure 1. Item Characteristic Curves of the Mathematics DIF

Items – French Immersion versus Francophone ... 56

Figure 2. Item Characteristic Curves of the Social Studies DIF

Items – French Immersion versus Francophone ... 57

Figure 3. Item Characteristic Curves of the Mathematics DIF

Items – FIE versus FIF 58

Figure 4. Item Characteristic Curves of the Social Studies DIF

Items – FIE versus FIF 59

A Complete Test of Hulin's Psychometric Theory of Measurement Equivalence
on Translated Tests

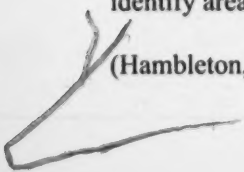
Large-scale achievement testing is pervasive in Canada, with nine out of ten provinces having provincial achievement testing programs. In addition, the Council of Ministers of Education conducts national assessments in mathematics, science, and reading and writing (e.g., Council of Ministers of Education, Canada, 1996). Due to Canada's official policy of bilingualism, many of the achievement tests initially created in English are later translated into French. This practice is not limited to Canada. Increasingly, achievement tests are being translated for use in cross-cultural and cross-national research around the world. For example, the International Association for the Evaluation of Educational Achievement conducts international assessments such as the Third International Mathematics and Science Study. The Third International Mathematics and Science Study, conducted in 1995, included 60 participating countries representing 41 different languages (Hambleton, 1994). Translated tests are increasingly used to assess the educational knowledge and skills of individuals from different countries and of individuals who speak different languages (Allalouf, Hambleton, & Sireci, 1999). With these large-scale assessments comes the problem of ensuring that both the original and the translated versions of the test are equivalent.

The goal of test translation is "to maintain construct equivalence and content representation across the two language forms" (Allalouf et al., 1999, p. 185). Traditionally, the translation-back-translation method has been used to develop tests in other languages (Gierl, Rogers, & Klinger, 1999; Hambleton, 1993; Hambleton & Patsula, 1998; Hulin, 1987). For the translation-back-translation method, the test is

created in the source language and language specialists render the test into the target language. Then, other language specialists translate the test from the target language back to the source language (the back-translation version). The original source language version and the back-translation version are then compared to see how closely they match each other. The accuracy of a translated test is often evaluated using the judgmental method of comparing the two versions of the source language test.

However, the comparison between the two source language versions of the test may not generalize to the target language version of the test. The back-translators may translate the target version of the test such that it is similar to the original version of the test even though the target version is inaccurate (Hambleton, 1993; Hulin, 1987). If the back-translators are able to conceal a poor target version of the test, then non-equivalence between the original version of the test and the target version of the test will not be detected (Hambleton, 1993). Consequently, the psychometric equivalence of the tests can not be assured using only the translation-back-translation method (Allalouf et al., 1999; Hambleton, 1993; Hulin, 1987).

Unfortunately the assumption of item equivalence across languages is often made without the use of any statistical procedures to check the claim (van de Vijver & Leung, 1997). The International Test Commission supports the need for statistical procedures to ensure item equivalence across cultures and languages (Hambleton, 1994). The International Test Commission suggested that test developers use appropriate statistical techniques not only to establish item equivalence but also to identify areas of a test that may be inadequate for one or more of the intended groups (Hambleton, 1994). The International Test Commission also recommended that test



developers conduct differential item functioning (DIF) analyses to evaluate a test designed to be used in two or more cultural or language groups.

The accuracy of a translated test is crucial to ensure that both language versions of the test are fair to all examinees. In large-scale testing situations, DIF is a constant concern because poorly translated items may put some students at a disadvantage (e.g., Hambleton, 1994; Hambleton & Patsula, 1998; Principles for Fair student Assessment Practices in Canada, 1993). DIF occurs when two different groups of examinees have a different probability of answering an item correctly, after controlling for overall ability (Shepard, Camilli, & Averill, 1981). In the case of translated tests, DIF analyses allow test developers to compare the two language groups who wrote the same test. But with translated tests it is difficult to determine whether DIF is attributable to translation error or cultural differences between the two groups of examinees.

Little research has been done to explore the reasons for why translated test items function differentially across language groups (Allalouf et al., 1999; Gierl & Khaliq, 1999). Hulin (1987) put forth a method using item response theory (IRT) for evaluating translated tests that could help determine why an item is functioning differentially when two language groups are compared. Hulin suggested that the item characteristic curves of the two groups be compared. Discrepant item characteristic curves indicate nonequivalence between the two groups of examinees. Moreover, he noted that the magnitude and direction of the discrepancies could possibly indicate the reasons for the DIF items.

Specifically, Hulin (1987) proposed that the a parameter differences (i.e., item discrimination parameter) indicated cultural differences whereas the b parameter

differences (i.e., item difficulty parameter) indicated translation errors. That is, the a parameter indicates how the item is discriminating between examinees with different ability levels. Differences in the a parameter mean that the item discrimination is not the same for the two groups of examinees. Hulin suggested that culture would be the influencing factor in how examinees responded to the translated items. On the other hand, the b parameter indicates the difficulty of the item. Differences across two groups of examinees in item difficulty means that the item is harder for one group relative to the other group of examinees. Hulin believed that differences due to item difficulty were the result of problems in the translation of the test resulting in some items that were harder for some examinees. Hulin suggested that statistical differences in the a and b parameters may provide information about psychological differences between language groups. This link between statistics and psychology could be critical to test translation because it would allow test developers to interpret statistical results substantively.

Hulin (1987) also described two types of comparisons that can be made to test his hypothesis. The first comparison is between two different cultural groups who speak the same language (different culture, same language). An example would be to compare Hispanic Americans to Mexicans and having both groups write the test in Spanish. Hulin hypothesized that there would be more a parameter differences when comparing examinees from different cultures in the same language. The only difference in this comparison is the culture of the two groups because the testing language remains the same. Since culture is an influencing factor on examinees responses to test items, Hulin suggested that there would be differences in the a parameter for the items displaying

DIF. The other comparison suggested by Hulin is to compare the same cultural group but in two different languages (same culture, different language). For example, comparing Hispanic Americans with some examinees writing in English and other examinees writing the same test in Spanish. As well, Hulin hypothesized that there would be more μ parameter differences when comparing examinees of the same culture in different languages. For this second comparison, the culture of the examinees remains constant and it is the testing language that is changed. Hulin suggested that problems in the translation accuracy would make the items harder for one group of examinees compared to the other group of examinees. Therefore, DIF would be the result of item difficulty.

There are problems when trying to compare examinees with the same cultural background in two different languages. The first problem is trying to find bilingual examinees with the same cultural background. People who speak different languages also tend to have different cultural backgrounds (Allalouf et al., 1999). In Canada, however, there is the opportunity to find bilingual examinees because of the federal government's official policy of bilingualism and because students can enrol in bilingual educational programs. It is possible, therefore, to find examinees in Canada who speak the same language and who come from a variety of cultural backgrounds. Another problem is the ability of making generalizations based on research using bilingual examinees (Ellis, 1995; Hambleton, 1993). Bilingual examinees may differ from monolingual examinees in cognitive skills such as divergent thinking (Diaz, 1983, as cited in Ellis, 1995). The test scores of bilingual examinees may be higher than the test scores of monolingual examinees (Hambleton, 1993). Consequently, the results of

research involving bilingual examinees may not be generalizable to monolingual examinees.

Even though several researchers suggested that there are difficulties testing Hulin's theory (Allalouf et al., 1999; Ellis, 1995; Hambleton, 1993; Hambleton, 1994; Hambleton & Patsula, 1998), Hulin (1987) started to evaluate his theory by analyzing the Job Descriptive Index (JDI), a measurement of job satisfaction. The JDI was translated from English into Spanish, Tagalog, Hebrew, and Canadian French. Three thousand subjects from five different countries responded to the JDI in either the source language or in one of the target languages. The cultural groups that responded to the JDI were Hispanic Americans, Mexicans, Canadian Anglophones, Canadian Francophones, Filipinos, and Americans. Then, pairwise comparisons of the item characteristic curves were made. Hulin used a chi-square test to evaluate the differences between the groups. Some of the comparisons between groups involved the same language but different countries (e.g., Hispanic Americans versus Mexicans both responding in Spanish), different languages yet the same country (e.g., Canadian Francophones responding in French versus Canadian Francophones responding in English), or different languages and different countries (e.g., Americans responding in English versus Canadian Francophones responding in French). Hulin found several items that were not equivalent between groups. Hulin discussed the possible reasons for the discrepant item characteristic curves, such as differences in item discrimination (a parameter) or item difficulty (b parameter). Although Hulin provided some ideas about the causes of DIF, he did not map the discrepant items to differences in the a or b parameters. The analyses that he presented, therefore, were not complete, leaving his hypothesis untested.

Ellis (1995) presented a partial test of Hulin's hypothesis by varying the culture and holding the language constant – different culture, same language comparison. Ellis examined the personality differences of 300 East Germans and 298 West Germans two years after the fall of the Berlin wall using the Trier Personality Inventory. Only four of 118 items were identified as displaying DIF: Two items were significant and two approached statistical significance. Of these four DIF items, three had large differences in the b parameter with only modest differences in the a parameter. The results of Ellis' study failed to empirically support Hulin's hypothesis because it was expected that there would be more differences in the a parameter than in the b parameter.

There are some key limitations with Ellis' (1995) partial test of Hulin's (1987) theory. The first limitation is the number of instruments used. Ellis used only one instrument, the Trier Personality Inventory, and therefore she Hulin's hypothesis with only tested only one instrument. The second limitation is the small sample sizes – sample sizes of 300 and 298 were used. Ideally 1500 or more subjects are required when using the 3-parameter IRT model (Hulin, Lissak, & Drasgow, 1982). Small sample sizes adversely affect IRT parameter estimation and can cause inaccurate parameter estimates. Ellis noted that the estimation of the a parameter is unstable even with larger sample sizes. The third limitation is the small number of DIF items. Ellis only found two DIF items and two near DIF items out of a total of 118 items. However, with so few items displaying DIF from a large set of items (118 items), it is possible that these items might have been detected due to Type I error. Moreover, the small number of items is inadequate for identifying patterns or systematic outcomes in the data. As a result, there were not enough items to make solid conclusions or

interpretations about Hulin's psychometric theory of measurement equivalence. The last limitation with Ellis' study is the lack of an effect size measure. Although Ellis mentions what the a and b parameters were for each DIF item, she did not establish criteria or indicate whether the differences in the item parameters were large or small.

Consequently, a test of Hulin's (1987) hypothesis that overcomes these limitations is needed. Data available in Canada provide a good opportunity to test Hulin's psychometric theory of measurement equivalence. The translation of achievement tests in the Canadian province of Alberta provides the opportunity to compare students who enrol in French Immersion programs with students who enrol in Francophone programs (different culture, same language). In addition, a random sample of French Immersion students periodically write the achievement tests in English or French (same culture, different language) for equating purposes. Together, this design allows for a complete test Hulin's hypothesis.

The data also allow for improvements on the work of Ellis (1995). The first improvement is an increase in the number of instruments. Rather than using one test, data from achievement tests in the areas of mathematics and social studies will be used to evaluate Hulin's (1987) theory. The second improvement over Ellis' work is the ability to test both research designs suggested by Hulin resulting in a complete evaluation of his theory. Ellis partially tested Hulin's theory by comparing subjects from different cultures but who speak the same language. The other research design is to compare examinees from the same culture, but who speak different languages. Ellis suggested that her partial test of Hulin's theory should be improved with a study in which the culture remains the same and the language differs. The third improvement in

this study will be the use of effect size measures. The item parameters for the DIF items will be compared to determine whether the differences in the a and b parameters are large, moderate, or small using well established criteria consistent with current DIF research.

This study will use existing mathematics and social studies achievement test data from a large-scale testing program in Alberta. The mathematics and social studies achievement tests were used because the different language requirements needed for each subject. The mathematics test has a lower language demand than the social studies test (i.e., less reading for the mathematics test than the social studies test). The achievement tests, initially created in English, were translated into French. Two comparisons were made in the present study. The first comparison was between French Immersion and Francophone students who wrote the achievement tests in French: the different culture, same language comparison. The second comparison was between French Immersion students. Some of the French Immersion students wrote the achievement test in English, while others wrote the tests in French: the same culture, different language comparison. The student responses were analysed using the computer program IRTDIF (Kim & Cohen, 1991) to calculate Lord's chi-square (Lord, 1980). Lord's chi-square was used because it is an IRT approach to identifying items displaying DIF. Lord's chi-square tests both the a and b parameters taking into account the variance-covariance of the item parameters. Since Lord's chi-square simultaneously tests the a and b parameters, it is not known which item parameter caused the item to be identified as displaying DIF. Therefore, the item parameters were compared and then

effect sizes were used to determine whether the differences between the groups were small, moderate, or large.

It is hypothesised that an *a* parameter difference indicates diversity between cultural groups and a *b* parameter difference indicates translation error. If Hulin's (1987) hypothesis is correct, then more *a* parameter differences than *b* parameter differences are expected when comparing examinees from different cultural groups even though examinees are tested in the same language. On the other hand, more *b* parameter differences than *a* parameter differences are expected when examinees from the same cultural are tested in different languages.

Method

Subjects

Two samples were examined in this study. The first sample compared French Immersion to Francophone students for the different culture, same language hypothesis. The second sample involved French Immersion students only. For this sample randomly selected students wrote the test in English and the remaining students wrote the test in French. In the Canadian province of Alberta, students in the French Immersion and Francophone programs take the same curriculum. However, there are two main differences between the French Immersion and Francophone programs. There are two different programs to meet the differing needs of the students. The French Immersion program was designed for students who are not native French speakers whereas the Francophone program was designed for native French speakers. The objective of the French Immersion program is to provide students with full mastery of English and functional fluency in French with an understanding and appreciation of the French

culture (Alberta Education, 1996). The objective of the Francophone program is to provide students with full master of French as the first language, full fluency of English, and an identity and belonging with the French community (Alberta Education, 1996). The Francophone students have differing educational, cultural, linguistic, and identity needs from the French Immersion students. As well, for the French Immersion program 50% of instruction time is in French and the other 50% of instruction time is in English. Instruction time for the Francophone program is all in French, with one exception. English is used for course in English. Not only are classes taught in French for the Francophone students, the school activities are also conducted in French. School programs are designed to give students the sense of belonging to the French culture and are encouraged to understand and participate in the French culture and community (Alberta Education, 1996). Whereas the French Immersion students are Anglophones learning French, the Francophone students learning English and French while trying to maintain the French culture. The Francophone students are considered to be culturally different from the French Immersion students.

The first comparison was made between the French Immersion students and Francophone students. There were 2200 French Immersion students for both the Mathematics and Social Studies Achievement Tests. There were 286 and 283 Francophone students for the Mathematics and Social Studies Achievement Tests, respectively. The French Immersion and Francophone students wrote the same 1997 Grade 6 Mathematics and Social Studies Achievement Tests that had been translated from English to French. There were 50 multiple-choice items on the mathematics test

and 49 multiple-choice items on the social studies test (these instruments will be described in the next section).

The second comparison was made between French Immersion students who wrote the achievement tests in English (FIE) and French Immersion students who wrote the same tests in French (FIF). There were 165 and 202 FIE students for the Mathematics and Social Studies Achievement Tests, respectively, and 178 FIF students wrote the Mathematics Achievement Test and 213 FIF students wrote the Social Studies Test. The FIE and the FIF students were randomly assigned, for equating purposes, to write the achievement tests in either English or French. Although the students wrote the achievement tests in different languages, they wrote the same 1995 achievement tests. There was 50 multiple-choice items on both the mathematics and social studies tests.

Instrument

The Mathematics Achievement Tests in both 1995 and 1997 contained 50 multiple-choice items, each item having four options. Each mathematics item was referenced to one of five content areas: numeration, operations and properties, measurement, geometry, and graphing. The items were also classified by cognitive level: knowledge or skills. The Social Studies Achievement Tests in 1995 and 1997 originally contained 50 multiple-choice items with each item having four options. One item was dropped from the 1997 Social Studies Achievement Test by Alberta Learning and was not included in the analyses. Each social studies item was referenced to one of four content areas: local government, Greece, China, and mapping and geography. Like the mathematics items, the social studies items were classified by cognitive level: knowledge or skills.

Mathematics and social studies were used because of the different language requirements needed for each academic subject. Social studies has a greater language demand compared to mathematics (i.e., the social studies tests have more words and require more reading than the mathematics tests). Mathematics was chosen because the math concepts are the same in both languages and the language demand in both English and French was small compared to social studies. These two academic subjects were used to ensure that the results of this study were not an artifact of the subject matter or its language demand.

Analyses

Construct equivalence. Student responses were fit to confirmatory factor analytic models to determine whether the test items were construct equivalent across the comparison groups (Gierl, 1999; van de Vijver & Poortinga, 1997). Construct equivalence is important because the items should be measuring the same construct regardless of the testing language. If the instrument is not measuring the same construct across different cultures or languages, then any conclusions based upon the instrument may be invalid (Frederiksen, 1977; Hambleton, 1993). Because differences between groups is the main concern of the present study, it is beneficial to know that any DIF item discovered is due to differences in the performance of the students and not caused by construct differences.

Three different models were tested to evaluate the construct equivalence across the groups of interest. The first model compared the factors to determine whether they were invariant across the groups. The second model tested the invariance of the factors and the factor loadings for the groups being compared. The third model tested the

invariance of the factors, the factor loadings, and the error variances. For this study, a one-factor model was tested. Three fit indices were used to assess each model: chi-square, root mean square error of approximation (RMSEA), and adjusted goodness-of-fit index (AGFI). A non-significant chi-square test indicates adequate model fit, as a non-significant model indicates that there are no differences between the models across the two groups. Note that the chi-square test is sensitive to sample sizes, therefore, Hayduk (1987) suggests that the chi-square test not be the only test to determine model-data fit. Browne and Cudeck (1993) suggested that a RMSEA value of 0.05 or less indicated a close model fit in relation to the degrees of freedom. Hayduk (1996) suggests an AGFI be at least 0.95 before the model has a chance of displaying no signs of ill fit. All confirmatory analyses were conducted with LISREL 8.14 (Jöreskog & Sörbom, 1993) using maximum likelihood estimation.

To test the three models, item parcels were created by summing items from the same content area. There are several advantages of using item parcels rather than using individual test items (Bandalos, 2000). The first advantage is the increase in reliability as item parcels are more reliable than individual items. As well, item parcels are continuous and normally distributed – a main assumption for maximum likelihood estimation. Another advantage is the reduction in the number of item parameters that LISREL will have to estimate. In other words, with item parcels there are fewer factor loadings and error variances that need to be estimated. Some researchers have also suggested that item parcels may be beneficial when sample sizes are small (Bandalos, 2000). For this study, the item parcels were based on the blueprint for each achievement test. As stated earlier there were five content areas for the mathematics tests:

numeration, operations and properties, measurement, geometry, and graphing. There are four content areas for the social studies tests: local government, Greece, China, and mapping and geography. To create the item parcels, students' scored responses were summed for each subscale.

Item parameter estimation. Next, with the computer program BILOG 3.11 (Mislevy & Beck, 1993), both the 2-parameter and the 3-parameter item response theory models were used to estimate the item parameters. Several different methods were used to determine whether the 2- or the 3-parameter IRT model better fit the data since either model could be used for this study. Because the assumption of guessing is different between the two models, the p-values for the ten hardest items on each test for the lowest scoring examinees were evaluated. If the p-values for the low scoring examinees were close to zero then the 2-parameter model would be more appropriate. However, if the p-values were approximately $1/(k-1)$, where k is the number of options, then the 3-parameter IRT model would be more appropriate. In addition to the p-values, the number of significant chi-square tests for the 2- and the 3-parameter models were compared. BILOG compares the parametric and non-parametric item characteristic curves to determine whether the model fits the data. A non-significant chi-square test indicates that the model-data fit is good.

After the item parameters have been estimated using BILOG, the item parameters had to be rescaled. Because the item parameters for each group were estimated separately, the item parameters for each group are on different metrics. Linear rescaling was used to establish a common score scale for the two groups of examinees in each of the two comparisons.

DIF analyses. The computer program IRTDIF (Kim & Cohen, 1991) identifies items that may be functioning differentially between the comparison groups. IRTDIF uses the estimated item parameters from BILOG to determine which items are functioning differentially. Lord's chi-square (Lord, 1980) was used to statistically identify which items functioned differentially. Lord's chi-square simultaneously tests the a and b parameters, taking into account the variance-covariance of the item parameters. The c parameter (the pseudo-chance parameter) is not considered in the analyses because the c parameter is constrained to the same value for both groups of examinees. The formula for Lord's chi-square is

$$\chi^2_{(2)} = (\xi_1 - \xi_2)' \Sigma^{-1} (\xi_1 - \xi_2)$$

where ξ are the estimates for the a and b parameters and Σ is the 2x2 dispersion matrix of the variance-covariance matrices. Therefore, an item is identified as displaying DIF when the item parameters are not identical within the limitations of sampling fluctuations (Hulin, Drasgow, & Parsons, 1983). Since the a and b parameters are being compared the chi-square test had two degrees of freedom and the critical value at $\alpha=0.05$ is 5.99.

Finally, of those items displaying DIF according to Lord's chi-square (1980), the a and b parameters were compared. The a and b parameter differences were classified as small, moderate, or large depending upon the magnitude of the difference. The magnitude of the item parameter differences were obtained from simulation research on uniform and nonuniform DIF. For simulation research, researchers have often used IRT to create the response vectors for each simulated examinee (Narayanan & Swaminathan, 1994; Narayanan & Swaminathan, 1996; Rogers & Swaminathan,

1993). By varying the magnitude of the difference in the a parameter or the b parameter, researchers were able to create small, moderate, or large amounts of DIF. A small amount of DIF occurred when the DIF effect size was less than 0.20 and a large amount of DIF occurred when the DIF effect size was greater than 0.80 (Narayanan & Swaminathan, 1994; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993).

In these simulation studies, the small effect size of .40 was associated with differences in the a parameter of 0.15 or less and differences in the b parameter of 0.50 or less. A large DIF effect size was associated with differences in the a parameter of 0.40 or larger and differences in the b parameter of 0.80 or larger. Therefore, an a parameter difference of 0.15 or less was considered to be a small difference. A difference in the a parameter between 0.15 and 0.40 was considered to be a moderate difference. Any a parameter difference over 0.40 was considered to be a large difference. A b parameter difference of 0.50 or less was considered to be a small difference. A moderate b parameter difference was between 0.50 and 0.80. A large b parameter difference was over 0.80 (Narayanan & Swaminathan, 1994; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). The a and b parameters have different values for determining what is small, moderate, or large DIF because the a and b parameters have different scales. The scale for the a parameter theoretically is from negative infinity to positive infinity. However, since negatively discriminating items are considered "poor items", the scale for the a parameter starts from zero and goes to positive infinity, with most a parameter values falling between zero and two (Hambleton et al., 1991, p.15). The scale for the b parameter is from negative infinity to positive infinity, with most b parameters falling between negative two and two (Hambleton et al., 1991, p. 13). As the

scales for the two item parameters are different, the a and b parameters have differing values for determining a small, moderate, or large parameter difference. Taken together, Lord's chi-square test and the parameter effect size measures were used to classify DIF items in this study.

Results

The results are presented in two parts. The results for the different culture, same language comparison are provided in the first part and the second part for the same culture, different language comparison.

Different Culture, Same Language

Psychometric characteristics of the tests. The psychometric characteristics of the Grade 6 mathematics and social studies tests for the French Immersion and the Francophone students are presented in Table 1. For mathematics, the French Immersion students performed better than the Francophone students. The mean score for the French Immersion students was 37.12, $SD = 7.57$ and the mean of the Francophone students was 35.24, $SD = 8.01$. The distributions of the French Immersion and Francophone students were similar as the standard deviations were almost equal. The distribution of the French Immersion students was slightly more negatively skewed than the distribution of the Francophone students. The distribution of the French Immersion students was slightly more peaked than the distribution of the Francophone students. The internal consistency was the same for the French Immersion students and the Francophone students. The mathematics items were of a comparable difficulty level for the Francophone and French Immersion students because the means, standard deviations and range for item difficulty were very similar. The mathematics items

discriminated equally well for the French Immersion and the Francophone students because the means, standard deviations, and range for item discrimination were similar.

The French Immersion students also performed better than the Francophone students. The mean and standard deviation of the French Immersion students were 31.75 and 7.71, respectively, and the mean and standard deviation of the Francophone students were 29.93 and 7.84, respectively. The variation of the French Immersion students was equal to the variation of the Francophone students. The distribution of the French Immersion students was slightly more negatively skewed than the distribution of the Francophone students. The distribution of the French Immersion students was also more peaked than the distribution of the Francophone students. The internal consistency was the same for the French Immersion and Francophone students. The social studies items were equally difficult for the French Immersion and the Francophone students. The social studies items had similar means, standard deviations, and ranges for the French Immersion students and Francophone students (see Table 1). The social studies items were equally discriminating for the French Immersion and Francophone students. The item discrimination range was less for the French Immersion students, which was also reflected in the smaller standard deviation for item discrimination. Although there were slight differences between the French Immersion and Francophone students, the psychometric characteristics were similar for both the Mathematics and Social Studies Achievement Tests.

Construct equivalence. The construct equivalence between the French Immersion form and the Francophone form was tested using confirmatory factor analysis. Three models were tested to assess construct equivalence: a) equated factors,

b) equated factors and factor loadings, and c) equated factors, factor loadings, and error variances. The constructs were considered to be equivalent when the parameters (factors, factor loadings, and error variances) were invariant across the two forms. For the Mathematics Achievement Test, the chi-square tests of the three models were significant as shown in Table 2, the values of the RMSEA were all under 0.05 and the values of the AGFI were all above 0.95. Although the chi-square tests indicated poor model fit, the RMSEA and the AGFI tests, two fit statistics not sensitive to sample size, indicated good model fit. For the Social Studies Achievement Test,⁶ the chi-square tests were not significant, the values of the RMSEA were all under 0.05, and the values of the AGFI were all above 0.95. These results suggest that the data fit a one-factor model. Therefore, the Mathematics and Social Studies Achievement Tests are considered unidimensional.

The three models were then compared to determine whether there were differences between the models across groups. For both the Mathematics and Social Studies Achievement Tests, model 1 versus model 2 and model 2 versus model 3 did not differ. The mathematics and social studies test forms for the French Immersion and Francophone students have comparable factors, factor loadings, and error variances. Consequently, the Mathematics and Social Studies Achievement Tests are considered to be psychometrically equivalent across groups.

Parameter estimation. The computer program BILOG was used to estimate the parameters for the 2- and 3-parameter IRT model. The 3-parameter IRT model had better model-data fit than the 2-parameter IRT model. The model-data fit was assessed by the number of significant chi-square tests and the guessing parameter for the ten

hardest test items (for the lower 25% of the examinees). The 2-parameter IRT model has 17 significant chi-square tests for the French Immersion form of the mathematics test and five significant chi-square tests for the Francophone form. However, with the 3-parameter model there was only six and two significant chi-square tests for the French Immersion and Francophone forms, respectively.

In addition, of the ten hardest items on the mathematics test for the French Immersion students, seven items had p-values above 0.20. There were six items with p-values above 0.20 on the mathematics test for the Francophone students. The mean p-value for the ten hardest items for the French Immersion students was 0.32 and for the Francophone students was 0.27. The lower 25% of the French Immersion and the Francophone students appear to be guessing at the ten hardest items on the test. The 3-parameter model is the better model for the Mathematics Achievement Test because the lower 25% of the students are able to obtain the correct response and there are few significant chi-square tests with the 3-parameter IRT model.

Similarly, the 3-parameter IRT model is more appropriate for the Social Studies Achievement Test as indicated by the number of significant chi-square tests and by the p-values. Compared to the 3-parameter IRT model there were more significant chi-square tests for the 2-parameter IRT model for both the French Immersion (14 versus 4) and the Francophone forms (9 versus 3).

As well, the lowest 25% of the students from both the French Immersion and the Francophone groups were correctly guessing at the ten hardest items. With the 2-parameter model, the ten hardest items for both the French Immersion students and the Francophone students had p-values all above 0.20. With the 3-parameter model, all of

the ten hardest items had p-values above 0.20. Therefore, the lowest 25% of the students on the Social Studies Achievement Tests were successfully obtaining the correct response for the hardest items close to the level of chance. The 3-parameter IRT model had the best model-data fit for the Mathematics and Social Studies Achievement Tests.

DIF analyses. When comparing the same language, different culture it was anticipated that there would be more a parameter differences than b parameter differences. Lord's chi-square (with the c-parameter constrained at 0.20¹) identified three DIF items in mathematics (items 1, 5, and 28, see Table 3.). Item 1 was associated with a large a parameter difference and a moderate b parameter difference. Item 5 was associated with a moderate a parameter difference and no difference in the b parameter. Item 28 was associated with a large a parameter difference and a moderate b parameter difference.

These differences are best illustrated graphically. Figure 1 presents the item characteristic curves for the three DIF items. For item 1, the French Immersion students were favoured until about 0 on the theta scale. Above 0 on the theta scale this first item favoured the Francophone students. Item 1 is a nonuniform DIF item because the item characteristic curves cross each other at about 0 on the theta scale. Item 5 is also a nonuniform DIF item. The French Immersion students are favoured from -4 to -2 on the theta scale. For the average and high ability levels the Francophone group is favoured. For the most part item 28 favoured the French Immersion students. The Francophone students are only very slightly favoured between 1 and 2 on the theta scale.

For the Social Studies Achievement Test, Lord's chi-square test indicated that there were five DIF items (items 2, 29, 39, 40, and 41; see Table 4). Item 2 had a moderate a parameter difference and no difference in the b parameter. Item 29 had a large a parameter difference and a moderate b parameter difference. Items 39, 40, and 41 all had large a -parameter differences, and no differences in the b parameter.

The social studies items are displayed graphically in Figure 2. As shown all items are nonuniform DIF items. For item 2, the French Immersion students are favoured below -1 on the theta scale and above -1 the Francophone students are favoured. Item 29 had the French Immersion students favoured from -4 to -0.5 on the theta scale, then above -0.5 the Francophone students are favoured for item 29. For item 39, the French Immersion students are favoured from -4 to -0.8 on the theta scale, then above -0.8 on the theta scale the Francophone students are favoured. For item 40, the French Immersion students are favoured for the lower part of the theta scale, below -0.5 on the theta scale, and then the Francophone students are favoured. For item 41, the French Immersion students are favoured below -0.9 on the theta scale. The Francophone students are favoured between -0.9 and 2.5. Above 2.5 the French Immersion students and the Francophone students have about the same probability of obtaining the correct response as do the Francophone students.

In summary, the two test forms had similar psychometric characteristics. Also, both tests were unidimensional as they fit a one-factor model. The test items were also construct equivalent because the two test forms had comparable factors, factor loadings,

¹ IRTDIF cannot calculate Lord's chi-square unless the c -parameter is constrained to the same value for both groups of examinees.

and error variances. The constrained 3-parameter logistic model was used for Lord's chi-square (1980) test. Eight items were identified as functioning differentially between the French Immersion students and the Francophone students. Of these eight items, all eight had large or moderate differences in the a parameter with three items also having large or moderate differences in the b parameter.

Same Culture, Different Language

Psychometric characteristics of the tests. The psychometric characteristics of the Grade 6 mathematics and social studies tests for the French Immersion students who wrote in English (FIE) and French Immersion the students who wrote in French (FIF) are presented in Table 5. For mathematics, the FIE students did better than the FIF students. For mathematics, the means were 38.49 and 37.36 for the FIE and FIF students, respectively. The FIE and the FIF students had similar variation as the standard deviations were essentially equal (7.40 and 7.49 respectively). The distribution of the FIE students was more negatively skewed than the distribution of the FIF students. Also, the distribution of the FIE students was slightly more peaked than the distribution of the FIF students. The internal consistencies were comparable between the FIE and the FIF students. The mathematics items were equal in difficulty for the FIE and the FIF students because the means, standard deviations, and item difficulty ranges were alike. The mathematics items were alike in the discrimination of the FIE and FIF students, even though the range for the FIE students was slightly larger than the range for the FIF students.

For the Social Studies Achievement Test, the FIE students outperformed the FIF students. The mean of the FIE students was 33.75, while the mean and standard

deviation of the FIF students was 29.23. The variability of the FIE students ($SD = 8.03$) was almost identical to the variability of the FIF students ($SD=8.13$). The distribution of the FIE students was more negatively skewed than the distribution of the FIF students. The distribution of the FIE students was also more peaked than the distribution of the FIF students. The internal consistency for the two groups was alike. The social studies items were on average easier for the FIE students than the FIF students ($M = 0.76$, $SD = 0.14$ and $M = 0.58$, $SD = 0.12$, respectively). Although the upper end of the item difficulty range was the same between the FIE and the FIF students, the lower end of the range was higher for the FIE students. The difference in the range for item difficulty suggests that some items were more difficult for the FIF students. The social studies items had similar item discrimination for the FIE and the FIF students. The mean, standard deviation, and range for item discrimination were similar between the two groups.

Construct equivalence. The construct equivalence between the French Immersion forms of the Mathematics and Social Studies Achievement Tests were tested using confirmatory factor analysis. Like, the different culture, same language hypothesis, three models were tested: a) equated factors, b) equated factors and factor loadings, and c) equated factors, factor loadings, and error variances. For the Mathematics Achievement Test, all tests suggested that there is adequate model fit (see Table 6). The chi-square tests were nonsignificant, the RMSEA values were all lower than the recommended 0.05, and the AGFI values were all above 0.95. The first model did not significantly differ from the second model, and the second model did not differ from the third model. The test forms had comparable factors, factor loadings, and error

variances. Therefore, there was construct equivalence on the Mathematics Achievement Test between the FIE and the FIF forms. Not only were the two test forms construct equivalent, the test forms were also unidimensional.

There is also adequate model fit for the Social Studies Achievement Test. The chi-square tests were all nonsignificant and the values of the AGFI were all 1.00. There was no value for the RMSEA because the values of the chi-square test were larger than the degrees of freedom. A negative value occurred when the degrees of freedom were subtracted from the value of the chi-square test [i.e., for the first model $\chi^2(7) = 3.99$, $3.99 - 7 = -3.01$]. Like the Mathematics Achievement Test, the social studies test had comparable factors, factor loadings, and error variances across the two test forms. The Social Studies Achievement Test was unidimensional because the test forms fit the one-factor model

Parameter estimation. Parameters for the 2- and 3-parameter IRT model were estimated using the computer program BILOG. For mathematics, the 3-parameter IRT model had better model-data fit than the 2-parameter model. Although the FIE form had only two significant chi-square with the 2-parameter model, the FIE form had 3 significant chi-square tests with the 3-parameter model. The FIF form, however, had fewer significant chi-square tests with the 2-parameter model compared to the 3-parameter model, 7 and 4 significant chi-square tests respectively.

In addition, the p-values for the lowest 25% of the examinees also indicated that the 3-parameter IRT model provided better model-data fit. The average p-values for the 10 hardest items were above 0.35 for both the FIE and the FIF forms of the Mathematics Achievement Test. Nine out of 10 items had p-values above 0.20 for the

lowest 25% of the examinees. The low ability examinees are able to correctly answer the ten hardest items close to the level of chance.

Similarly, the 3-parameter model IRT model is also more appropriate for the Social Studies Achievement Test. The number of significant chi-square tests dropped from six to zero, for the 2- and 3-parameter model respectively, for both the FIE and the FIF forms of the Social Studies Achievement Test. As well, it appears the lowest 25% of the students were able to correctly guess the answers on the ten hardest items. The p-values for the ten hardest items averaged over 0.25. The 3-parameter model provided better fit to the data because there was no significant chi-square tests and the lower ability students were able to correctly guess the answers to the ten hardest items. BILOG was not able to estimate the b parameter for one item on the FIF form of the social studies test. This item was dropped from the rest of the analyses. The 3-parameter IRT model was used because it best fit the data for mathematics and social studies.

DIF analyses. It was anticipated that there would be more b parameter differences than there would be a parameter differences for the different language, same culture hypothesis.

Lord's chi-square test, with the c-parameter constrained to 0.20 for both groups, identified three mathematics items with DIF (items 6, 17, and 38). The item parameter differences are reported in Table 7. Item 6 had no difference in the a parameter and a large b parameter difference. Item 17 was associated with a large a parameter difference and a moderate b parameter difference. Similarly to item 17, item 38 had large a parameter difference with a large b parameter difference.

Figure 3 has the item characteristic curves for items 6, 17, and 38. Item 6 is a uniform DIF item. The FIF students have a higher probability of obtaining the correct response across the entire theta scale. Item 17 is a nonuniform DIF item. The FIE students are favoured above -1.0 on the theta scale. Below -1.0 the FIF students are favoured. Item 38 is also a nonuniform DIF item. The FIE students are slightly favoured above 0.9 on the theta scale. From -4 to 0.9 on the theta scale the FIF students are greatly favoured over the FIE students.

For the Social Studies Achievement Test, Lord's chi-square test indicated that there were five DIF items (items 24, 25, 27, 35, and 48). Table 8 has the item parameter differences for the Social Studies Achievement Test. Item 24 had a large a parameter difference with no difference in the b parameter. Although item 25 was not associated with an a parameter difference, it had a large b parameter difference. Item 27 also had a large a parameter difference with a large b parameter difference. Item 35 had a large difference in the a parameter and was also associated with a moderate difference in the b parameter. Item 48 did not differ in the a parameter, but had a moderate b parameter difference.

These items are illustrated graphically in Figure 4. Item 24 is a nonuniform DIF item. The FIE students are favoured above -0.9 on the theta scale and below -0.9 on the theta scale the FIF students are favoured. Item 25 is a uniform DIF item. The FIF students have a higher probability of obtaining the correct response. For item 27, the FIE students are favoured from -4 to 0.9 on the theta scale. The FIF students are then favoured between 0.9 and 3. The FIE and the FIF students have the same probability of obtaining the correct response above 3. For the most part of item 35 the FIE students are

favoured. However, from above 1.2 on the theta scale the FIE and the FIF students have about the same probability of obtaining the correct response for item 35. Item 48 is the opposite of item 35 as the FIF students tend to be favoured. Below -2.5 and above 3 the FIE and the FIF students have about the same probability of obtaining the correct response for item 48.

In summary, the two test forms had similar psychometric characteristics. The test items were determined to be construct equivalent because the two test forms had comparable factors, factor loadings, and error variances. The Mathematics and Social Studies Achievement Tests were unidimensional because they fit the one-factor model. Using the constrained 3-parameter logistic model for Lord's chi-square test, eight items were identified as functioning differentially between the French Immersion students who wrote in English and the French Immersion students who wrote in French. Seven of the eight DIF items had large or moderate differences in the b parameter with five of the items having large or moderate differences in the a parameter.

Summary and Conclusion

Two comparisons were made in this study. The first comparison was between the French Immersion students and the Francophone students – different culture, same language comparison. The second comparison was between French Immersion students, some students wrote the achievement tests in English and others wrote the achievement tests in French – same culture, different language comparison. Following Hulin's (1987) theory, it was predicted that for the DIF items there would be more a parameter differences than b parameter differences when comparing the French Immersion students to the Francophone students. Further, it was predicted that for the DIF items

there would be $\underline{b}$ parameter differences than $\underline{a}$ parameter differences when comparing the French Immersion students (FIE vs FIF).

When comparing the French Immersion and the Francophone students Lord's chi-square identified eight items identified as displaying DIF (three mathematics items and five social studies items), all eight items had a large or a moderate $\underline{a}$ parameter difference with three items also having a moderate $\underline{b}$ parameter difference. This trend in the data follows Hulin's (1987) suggestion that there would be more $\underline{a}$ parameter differences than $\underline{b}$ parameter differences when comparing examinees from different cultures yet testing them in the same language.

Similarly to the comparison between French Immersion and Francophone students, a total of eight items were identified by Lord's chi-square as displaying DIF when the FIE and FIF students were compared (three items in mathematics and five items in social studies). Of these eight items, seven items had a large or a moderate $\underline{b}$ parameter difference with five of the items also having a large $\underline{a}$ parameter difference. As there are more $\underline{b}$ parameter differences than $\underline{a}$ parameter differences on the items identified as displaying DIF, the trend in the data conforms with Hulin's (1987) theory.

Surprisingly, 11 out of the 16 DIF items were nonuniform DIF and the rest of the items were uniform DIF items. The difference between uniform and nonuniform DIF is the crossing of the item characteristic curves. For a uniform DIF item, the item characteristic curves do not cross over whereas for a nonuniform DIF item the item characteristic curves cross (Camilli & Shepard, 1994). When the item characteristic curves cross there is an interaction between ability level and group membership (Narayanan & Swaminathan, 1996). This is a surprising result because many

researchers have noted that nonuniform DIF is not as common as uniform DIF (Camilli & Shepard, 1994; Gierl et al., 1999; Mazor, Clauser, & Hambleton, 1994; Narayanan & Swaminathan, 1996). In this study the high number of nonuniform DIF items may be the result of the a and b parameter differences. Nonuniform DIF occurs because of the interaction between the a and the b parameters. This outcome suggests that there is an interaction between culture and translation accuracy. Therefore, both cultural variation and translation differences need to be considered simultaneously when trying to understand DIF. The occurrence of nonuniform DIF may also be caused by poor parameter estimation. The 3-parameter model requires large sample sizes for accurate and stable parameter estimates. The sample sizes used in this study were small and therefore the item parameters may be inaccurate. With larger sample sizes, it is possible that the nonuniform DIF items might instead be uniform DIF items. But further research is needed to resolve the nature of this outcome.

The trend in the data from both the different culture, same language and the same culture, different language comparisons conforms with Hulin's (1987) psychometric theory of measurement equivalence on translated tests. These results also contradict the previous findings by Ellis (1995) who partially tested Hulin's theory. Ellis' findings did not support Hulin's theory because she found more b parameter differences rather than a parameter differences. Recall that Ellis compared East Germans to West Germans on the Trier Personality Inventory – different culture, same language comparison – therefore, she expected more a parameter differences than b parameter differences.

Limitations

Although the trends in the data follow what Hulin (1987) suggested, there are several limitations. The first limitation is the small sample sizes used in this study. There is only a small number of Francophone students within the Province of Alberta providing a population of less than 300 examinees. Therefore, even with the entire population of Francophone students, sample size is a limitation. As well, only a small number of French Immersion examinees wrote the Mathematics and Social Studies Achievement Tests in English.

The second limitation of this study is the unstable a parameter estimates. Small sample sizes affect the parameter estimation procedures. The a parameter tends to be an unstable item parameter. The small sample sizes makes the a parameter even more unstable. In this study the a parameter for item 27 was estimated to be 0.650 for the FIE students and 5.203 for the FIF students. It is possible that the difference of -4.553 may actually be much smaller. Larger sample sizes would have allowed for more stable estimation of the item parameters which would be reflected in the number of a parameter differences. It is possible that for the same culture, different language comparison the number of a parameters might have been actually less than was actually found. Any future tests of Hulin's psychometric theory should be conducted with larger sample sizes.

The third limitation is the use of bilingual examinees. The results of this study may not be generalizable to monolingual examinees. Some researchers suggest that bilingual examinees may differ from monolingual examinees for some cognitive abilities such as divergent thinking and other cognitive skills (Diaz, 1983, as cited in

Hambleton, 1993). Hambleton (1993) suggested that researchers should assess the degree of examinees' proficiency of each language. Unfortunately, it is very difficult to assess examinees' bilingualism (Ellis, 1995). Cognitive differences between monolingual and bilingual examinees may make it difficult to generalize the results of this study.

The last limitation is the use of Lord's chi-square. There is the concern that Lord's chi-square may not be an appropriate method for comparing item characteristic curves (Linn, Levine, Hastings, & Wardrop, 1981). Because Lord's chi-square tests the item parameters rather than the area between the item characteristic curves, it is possible that an item may be identified as displaying DIF when there is no practical difference between the item characteristic curves (Hambleton & Swaminathan, 1985; Hulin et al., 1983). For example, an item with the parameters of $a = 1.8$, $b = 3.5$, and $c = 0.2$ for one group and $a = 0.5$, $b = 5.0$, and $c = 0.20$ would be identified as displaying DIF by Lord's chi-square. However, there is no practical difference between the item characteristic curves for the theta interval of -3 to +3, the difference between the item characteristic curves is 0.05 (Hambleton & Swaminathan, 1985; Hulin et al., 1983). Another problem with Lord's chi-square is that it may not follow the chi-square distribution when the item and ability parameters are estimated at the same time (McLaughlin & Drasgow, 1987). Another disadvantage of using Lord's chi-square is the possibility of inflated type I errors due to underestimated standard errors (McLaughlin & Drasgow, 1987). Recall that Lord's chi-square takes into account the variance-covariance of the item parameters. With small standard errors, an item may be identified as displaying DIF when the item may actually be a non-DIF item.

Improvements

Even though there are some limitations to this study, there are also some noteworthy improvements over Ellis' (1995) research that may account for the different findings. The first improvement was an increase in the number of instruments tested. Two instruments were used in this study: a mathematics and a social studies test. The second improvement was the two comparisons. Ellis only compared West Germans and East Germans and tested all the subjects in German (different culture, same language comparison), which was a partial test of Hulin's psychometric theory. This study had the different culture, same language comparison and the same culture, different language comparison. Consequently, this study was a full test of Hulin's theory. The third improvement was an increase in the number of DIF items resulting in a better test of Hulin's theory. The fourth improvement over Ellis' study is use of an effect size measure for the differences in the a and b parameters. Lord's chi-square does not indicate whether differences in the item characteristic curves are a result of differences in the a or the b parameters. Criteria for parameter differences from current DIF research were used to determine whether the a and b parameter differences were large, moderate, or small.

Implications

Hulin (1987) suggested that statistics, specifically differences in the a and b item parameter, could provide test developers and researchers information about group differences on instruments that had been translated or adapted for use in other languages. The common translation-back-translation method for adapting instruments for use in other languages is problematic because the original source language version

of the instrument is compared to the final source language version. The source language version of the instrument is never compared with the target language version. Therefore, translation errors may not be discovered allowing for the possibility of incorrect and invalid interpretations to occur (Hambleton, 1993; Hambleton, 1994; Hambleton & Patsula, 1998). The translation-back-translation method requires a judgmental review to determine if the two source versions are similar. The International Test Commission (Hambleton, 1994) suggests that an objective process such as Hulin's psychometric theory of measurement equivalence be used in the evaluation of translated instruments.

Hulin's (1987) theory is useful to test developers who are interested in using a instrument in more than one language. Hulin's theory would assist test developers reviewing the DIF results of a test. Often test developers will remove items that have been identified as displaying DIF because the cause of DIF are difficult to identify. Hulin's theory provides an approach for evaluating DIF items. By looking at the parameter differences of the DIF items, test developers may have an idea as to why an item may function differentially across groups of examinees. Test developers might be able to change items with a parameter differences to make them more culturally relevant for the examinees. As well, test developers might be able to re-translate items with b parameter differences. By knowing the reason for the DIF, test developers may be able to make the appropriate changes to an item and maintain the size of their item bank.

Hulin's (1987) psychometric theory of measurement equivalence on translated tests is also useful to cross-cultural researchers. Currently, researchers are interested in understanding the causes of DIF on tests that have been translated and adapted for use

in a variety of languages (Allalouf et al., 1999; Gierl & Khaliq, 2000) and how to change poorly translated test items (Allalouf, 2000). The Simultaneous Item Bias Test (SIBTEST) is a common statistical procedure used to identify items displaying DIF. However, SIBTEST can only identify items that have uniform DIF (i.e., the item characteristic curves do not cross). IRT and Hulin's (1987) theory could assist researchers in identifying items displaying both uniform and nonuniform DIF and providing the researchers with reasons why the items may be displaying DIF. Gierl and Khaliq (2000) found that translation differences did not always explain why DIF items were identified. Perhaps Hulin's theory can provide some insight into the nature of DIF for items with no translation errors. Hulin suggested that some items will display DIF because of cultural variation between groups of examinees.

Research in the area of translation DIF has evolved from determining which items display DIF to the causes of DIF and how to prevent DIF on translated items. Hulin's (1987) theory provides an explanation for why translated test items may be functioning differentially. It is very difficult for researchers to look at all the items on a test and determine which items are functioning differentially and why. Hulin's theory allows a researcher to narrow the focus from an entire test to a few items. For example, in this study the number of items has dropped from a total of 198 to just 16 items. In addition, Hulin's theory also provides a suggestion as to the causes of the DIF. The 16 items can be separated into two groups of eight items. According to Hulin's theory, the cause of the DIF for the first eight items is cultural variation and the cause of the DIF for the last eight items is translation error. The guidance that Hulin's theory provides may be very useful to researchers interested in the causes of DIF. Engelhard, Hansche,

and Rutledge (1990) found that item reviewers are unable to accurately predict which items function differentially. Researchers have also noted the difficulty of interpreting the causes of DIF (Engelhard et al., 1990). Hulin's theory may be a solution to the problem of identifying and correctly interpreting the causes of DIF.

In response to the difficulties that researchers have in identifying and interpreting the causes of DIF, Roussos and Stout (1996) suggest that substantive and statistical procedures should be used together to better understand the causes of DIF. Hulin's (1987) theory is a combination of the statistical approach by using IRT and the substantive approach by suggesting that the a parameter indicates cultural variation and the b parameter indicates translation errors. Hulin's theory is an attempt to link psychological and statistical outcomes using IRT. For this reason, Hulin's theory warrants study.

But further research into Hulin's (1987) theory is needed to determine whether it is accurate. Hulin's theory allows test developers to make substantive inferences about the test from statistical outcomes and, therefore, make inferences about the psychological processes of the test-taking process. However, statistical outcomes do not provide any information about what an examinee is thinking about while reading and answering the test items. It is only Hulin's speculation that differences in the a parameter indicate cultural differences and differences in the b parameter indicate translation errors. As part of further research into Hulin's theory, protocol analyses is needed to completely understand the psychological processes of the examinees. Protocol analyses can provide information about the cultural differences between the examinees and whether the cultural differences influence the examinees' interpretation

of the test item and subsequently, the examinees' answers. In addition to protocol analyses, research on the accuracy of the test translation is necessary. Having certified translators, familiar with the cultures of interest, review different language versions of a test can provide information on which items are inaccurately translated. The combination of protocol analyses and translator reviews should provide further insight into Hulin's theory.

References

- Alberta Education. (1996). Yes, you can help! A guide for French Immersion parents. Edmonton, Alberta: Author.
- Allalouf, A. (2000). Retaining translated verbal reasoning items by revising DIF items. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.
- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. Journal of Educational Measurement, 36, 185 – 198.
- Bandalos, D. (2000). The effect of parceling on structural equation models. Unpublished manuscript.
- Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp.136 - 162). Newbury Park, CA: Sage.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
- Council of Ministers of Education. (1996). SAIP 1996 report on science. Toronto, ON: Author.
- Ellis, B. B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. European Journal of Psychological Assessment, 11, 184 – 193.

Engelhard, G., Jr., Hansche, L., & Rutledge, E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3, 347-360.

Frederiksen, N. (1977). How to tell if a test measures the same thing in different cultures. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology (pp. 14 – 18). Amsterdam: Swets and Zeitlinger B.V.

Gierl, M. J. (1999). Construct equivalence on translated achievement tests. Manuscript submitted for publication.

Gierl, M. J., & Khaliq, S. N. (2000). Identifying sources of differential item functioning on translated achievement tests: A confirmatory analysis. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, PQ.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57 – 68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229 – 244.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. Social Indicators Research, 45, 153 – 171.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hayduk, L. A. (1987). Structural equation modeling with LISREL essential and advances. Baltimore, MA: John Hopkins University Press.

Hayduk, L. A. (1996). LISREL: Issues, debates, and strategies. Baltimore, MA: John Hopkins University Press.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations, Fidelity across languages. Journal of Cross - Cultural Psychology, 18, 115 - 142.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, 6, 249-260.

Jöreskog, K. G. & Sörbom, D. (1993). LISREL 8.14: A computer program for structural equation modeling. Chicago, IL: Scientific Software.

Kim, S., & Cohen, A. S. (1991). IRTDIF: A computer program for IRT differential item functioning analysis. University of Wisconsin-Madison.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173

- Lord, F. M. (1980). Applications of item response theory. Hillsdale, NJ: Erlbaum.
- Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, 54, 284-291.
- Mislevy, R. J., & Bock, R. D. (1993). BILOG 3.11: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.
- Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. Applied Psychological Measurement, 18, 315-328.
- Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. Applied Psychological Measurement, 20, 257-274.
- Principles for Fair Students Assessment Practices for Education in Canada (1993). University of Alberta, Edmonton, AB.
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.
- Roussos, L., & Stout, W. (1996). A Multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.
- Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Journal of Educational and Statistics, 6, 317 - 375.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards and integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13, 21-29.

Table 1

Descriptive Characteristics for French Immersion and Francophone

Characteristic	Mathematics		Social Studies	
	Immersion	Francophone	Immersion	Francophone
No. of Examinees	2200	286	2200	283
No. of Items	50	50	49	49
Mean	37.12	35.24	31.75	29.93
Standard Deviation	7.57	8.01	7.71	7.84
Skewness	-0.65	-0.35	-0.34	-0.09
Kurtosis	-0.11	-0.61	-0.49	-0.66
Internal Consistency ^a	0.87	0.87	0.84	0.84
Mean Item Difficulty	0.74	0.70	0.65	0.61
SD Item Difficulty	0.14	0.15	0.12	0.13
Range Item Difficulty	0.22 - 0.94	0.20 - 0.91	0.39 - 0.87	0.35 - 0.85
Mean Item Discrimination ^b	0.45	0.44	0.38	0.37
SD Item Discrimination	0.11	0.14	0.11	0.14
Range Item Discrimination	0.09 - 0.67	0.06 - 0.72	0.17 - 0.59	0.03 - 0.67

^aCronbach's alpha.^bBiserial correlation.

Table 2

Tests for Model Equivalence Between French Immersion and Francophone Students

Content Area	χ^2	df	RMSEA	AGFI
Mathematics				
Model 1	37.45*	14	0.037	0.979
Equated Factors				
Model 2	42.36*	19	0.031	0.976
Equated Factors				
Equated Factor Loadings				
Model 3	44.64*	20	0.031	0.978
Equated Factors				
Equated Factor Loadings				
Equated Error Variances				
Social Studies				
Model 1	9.81	7	0.018	0.956
Equated Factors				
Model 2	12.02	11	0.009	0.982
Equated Factors				
Equated Factor Loadings				
Model 3	12.29	12	0.004	0.983
Equated Factors				
Equated Factor Loadings				
Equated Error Variances				
<u>Model Comparison</u>	χ^2	df		
Mathematics				
Model 1 vs Model 2	4.91	5		
Model 2 vs Model 3	2.28	1		
Social Studies				
Model 1 vs Model 2	2.21	4		
Model 2 vs Model 3	0.27	1		

* $p < 0.01$

Table 3.

Item Parameters for the Mathematics Items – French Immersion versus Francophone

Item	a parameter			b parameter		
	FI	FR	Difference	FI	FR	Difference
1	0.621	1.727	-1.106	-1.298	-0.500	-0.798
2	0.483	0.371	0.111	-2.065	-2.867	0.802
3	0.633	1.037	-0.404	0.553	0.270	0.283
4	0.673	0.687	-0.014	0.470	0.227	0.243
5	0.515	0.746	-0.231	-1.181	-1.426	0.245
6	0.536	0.705	-0.168	-1.586	-1.289	-0.297
7	0.657	0.682	-0.025	-0.422	-0.646	0.224
8	0.448	0.609	-0.161	-1.758	-1.430	-0.328
9	0.437	0.379	0.058	-2.954	-3.567	0.613
10	0.592	0.656	-0.064	-1.035	-1.037	0.002
11	0.810	0.734	0.076	-0.246	-0.506	0.260
12	0.921	0.817	0.104	-0.267	-0.338	0.070
13	0.487	0.367	0.120	-1.609	-2.404	0.795
14	0.739	0.830	-0.091	-1.003	-0.955	-0.049
15	0.960	1.214	-0.254	-1.720	-1.426	-0.294
16	1.335	1.390	-0.055	2.145	1.657	0.488
17	0.922	0.620	0.302	-1.722	-1.999	0.277
18	0.764	0.778	-0.014	0.065	0.157	-0.092
19	0.820	0.673	0.147	-0.671	-0.618	-0.053
20	0.756	0.822	-0.067	-1.258	-1.144	-0.114
21	0.583	0.521	0.061	-1.131	-1.368	0.237
22	0.538	0.885	-0.347	-2.320	-1.646	-0.674
23	0.873	0.697	0.176	-1.789	-2.093	0.304
24	0.373	0.430	-0.057	-1.500	-1.540	0.040
25	0.558	0.537	0.020	-1.984	-1.987	0.003
26	0.484	0.767	-0.284	-1.869	-1.613	-0.256
27	0.887	0.866	0.022	-0.807	-1.257	0.449
28	0.894	1.481	-0.588	-0.830	-0.225	-0.605
29	0.389	0.571	-0.182	-1.594	-0.976	-0.619
30	0.555	0.567	-0.012	-1.974	-2.159	0.184
31	0.698	0.858	-0.160	-2.400	-1.964	-0.437
32	0.855	1.030	-0.175	0.236	0.181	0.055
33	0.592	0.763	-0.171	-0.760	-0.814	0.054
34	0.622	0.747	-0.126	-1.953	-1.726	-0.227
35	0.738	0.924	-0.186	-0.376	-0.273	-0.102
36	0.842	1.165	-0.322	-1.416	-1.450	0.034
37	1.172	1.259	-0.087	-0.535	-0.473	-0.061
38	0.757	1.022	-0.265	0.760	0.730	0.030
39	1.006	1.129	-0.123	0.254	0.450	-0.196

40	0.372	0.530	-0.158	-1.434	-0.907	-0.526
41	0.837	0.724	0.113	0.327	-0.108	0.435
42	0.618	0.591	0.027	0.023	-0.190	0.213
43	0.774	0.814	-0.040	-0.284	-0.240	-0.044
44	0.531	0.699	-0.168	-0.775	-0.840	0.065
45	0.579	0.539	0.039	0.186	0.206	-0.020
46	0.825	1.299	-0.474	-1.080	-1.016	-0.065
47	0.678	0.988	-0.310	-0.803	-0.671	-0.132
48	1.124	1.537	-0.413	0.176	0.347	-0.171
49	1.233	1.397	-0.164	-0.184	-0.015	-0.169
50	0.980	1.248	-0.268	-0.514	-0.435	-0.079

Note. A negative difference indicates that the parameter is larger for the Francophone students (French Immersion minus Francophone = difference).

Table 4.

Item Parameters for the Social Studies Items – French Immersion versus Francophone

Item	a parameter			b parameter		
	FI	FR	Difference	FI	FR	Difference
1	0.350	0.471	-0.121	-0.309	-0.662	0.353
2	-0.309	0.551	-0.243	-0.056	-0.526	0.470
3	0.457	0.356	0.100	-0.198	0.117	-0.315
4	0.376	0.450	-0.074	-1.556	-1.668	0.112
5	0.826	0.672	0.154	-0.451	-0.784	0.333
6	0.488	0.998	-0.510	-0.264	-0.318	0.054
7	0.947	0.886	0.061	-0.688	-1.217	0.530
8	0.481	0.638	-0.157	0.847	0.706	0.141
9	0.657	0.840	-0.183	-0.119	-0.202	0.084
10	0.602	0.717	-0.114	-0.279	-0.030	0.309
11	0.318	0.361	-0.043	-0.800	0.759	-1.559
12	0.710	0.860	-0.150	0.039	-0.218	0.257
13	0.678	0.932	-0.254	-0.580	-0.255	-0.325
14	1.031	0.750	0.281	-0.160	-0.352	0.192
15	0.542	0.701	-0.158	-0.700	-0.685	-0.015
16	0.366	0.674	-0.309	0.497	0.152	0.344
17	0.749	0.747	0.002	1.001	1.122	-0.120
18	0.424	0.868	-0.444	1.133	1.484	-0.350
19	0.709	0.890	-0.181	-0.443	-0.643	0.200
20	0.705	0.789	-0.084	-0.264	-0.244	-0.020
21	0.965	1.122	-0.157	-0.455	-0.628	0.173
22	0.608	0.816	-0.208	-0.368	-0.207	-0.161
23	0.257	0.909	-0.651	1.727	2.391	-0.664
24	0.898	1.411	-0.513	-0.179	-0.168	-0.011
25	0.764	1.062	-0.299	0.824	0.452	0.373
26	0.784	0.970	-0.185	0.104	0.297	-0.193
27	0.705	0.880	-0.175	-1.235	-1.116	-0.119
28	0.475	0.781	-0.305	-0.733	-0.339	-0.394
29	0.535	1.004	-0.469	0.851	0.311	0.541
30	0.430	0.654	-0.223	-1.392	-0.996	-0.396
31	0.413	0.563	-0.150	-0.889	-0.284	-0.605
32	0.515	0.844	-0.329	1.875	1.504	0.371
33	0.251	0.471	-0.220	-0.241	-0.047	-0.194
34	0.565	0.498	0.067	-1.031	-1.200	0.169
35	0.500	0.909	-0.410	1.195	1.208	-0.012
36	0.762	0.585	0.177	0.042	-0.188	0.229
37	0.613	0.747	-0.134	-1.588	-1.435	-0.153
38	0.932	1.173	-0.242	-1.351	-1.312	-0.039
39	0.792	1.393	-0.600	-0.361	-0.527	0.165
40	0.514	0.926	-0.412	0.440	0.131	0.309

41	0.833	1.468	-0.635	-0.479	-0.638	0.159
42	0.989	1.185	-0.196	0.354	0.207	0.147
43	0.638	0.802	-0.165	-0.100	-0.216	0.115
44	0.923	1.277	-0.354	-0.054	-0.235	0.182
45	0.928	1.328	-0.401	0.605	0.410	0.195
46	1.037	1.482	-0.445	0.637	0.683	-0.046
47	0.890	0.567	0.323	0.605	0.625	-0.020
48	0.545	1.015	-0.470	-0.961	-0.249	-0.712
49	0.742	0.913	-0.172	-1.603	-1.517	-0.086

Note. A negative difference indicates that the parameter is larger for the Francophone students (French Immersion minus Francophone = difference).

Table 5.
Descriptive Characteristics for French Immersion Students

Characteristic	Mathematics		Social Studies	
	English	French	English	French
No. of Examinees	165	178	202	213
No. of Items	50	50	50	50
Mean	38.49	37.36	33.75	29.23
Standard Deviation	7.40	7.49	8.13	8.03
Skewness	-0.82	-0.43	-0.50	-0.09
Kurtosis	-0.26	-0.73	-0.32	-0.67
Internal Consistency ^a	0.87	0.86	0.86	0.84
Mean Item Difficulty	0.77	0.75	0.76	0.58
SD Item Difficulty	0.13	0.13	0.14	0.12
Range Item Difficulty	0.32 - 0.98	0.36 - 0.99	0.37 - 0.87	0.27 - 0.85
Mean Item Discrimination ^b	0.48	0.44	0.43	0.37
SD Item Discrimination	0.16	0.17	0.14	0.14
Range Item Discrimination	0.00 - 0.87	0.05 - 0.75	0.03 - 0.67	0.01 - 0.62

^aCronbach's alpha.

^bBiserial correlation.

Table 6.

Tests for Model Equivalence Between French Immersion Students Who Wrote inEnglish and Who Wrote in French

Content Area	χ^2	df	RMSEA	AGFI
Mathematics				
Model 1	14.88	14	0.019	0.9786
Equated Factors				
Model 2	20.35	19	0.020	0.9763
Equated Factors				
Equated Factor Loadings				
Model 3	20.35	20	0.010	0.9775
Equated Factors				
Equated Factor Loadings				
Equated Error Variances				
Social Studies				
Model 1	3.99	7	0.00	1.00
Equated Factors				
Model 2	6.39	11	0.00	1.00
Equated Factors				
Equated Factor Loadings				
Model 3	6.41	12	0.00	1.00
Equated Factors				
Equated Factor Loadings				
Equated Error Variances				
<u>Model Comparison</u>	χ^2	df		
Mathematics				
Model 1 vs Model 2	5.47	5		
Model 2 vs Model 3	0	1		
Social Studies				
Model 1 vs Model 2	2.40	4		
Model 2 vs Model 3	0.02	1		

Note. The formula for RMSEA is $\chi^2 - df = x/n = x/df = \sqrt{x} = \sqrt{(F_0/d)} = \varepsilon_a$. For the three social studies models, the degrees of freedom are larger than the value of the χ^2 providing a negative value which can not be square rooted. Therefore, the values for RMSEA have been set to zero.

Table 7.

Item Parameters for the Mathematics Items – FIE versus FIF

Item	a parameter			b parameter		
	FIE	FIF	Difference	FIE	FIF	Difference
1	1.005	0.734	0.272	-2.645	-2.977	0.333
2	1.257	1.452	-0.195	-0.367	-0.435	0.068
3	1.164	0.935	0.229	-1.471	-2.356	0.885
4	0.751	0.533	0.218	-1.300	-1.279	-0.021
5	0.870	0.617	0.253	-1.911	-2.414	0.503
6	0.590	0.488	0.101	1.571	-1.034	2.605
7	0.511	0.422	0.089	-0.539	-0.214	-0.325
8	0.856	0.717	0.139	-2.010	-1.875	-0.135
9	0.797	0.738	0.059	-0.948	-1.065	0.118
10	0.751	0.454	0.298	-1.233	-0.914	-0.319
11	0.907	0.520	0.387	-0.926	-1.339	0.414
12	0.636	0.825	-0.189	0.255	0.513	-0.258
13	0.745	0.641	0.104	0.032	-0.127	0.160
14	0.852	1.153	-0.301	-0.423	-0.016	-0.407
15	0.850	1.019	-0.169	-0.439	-0.146	-0.293
16	0.882	0.692	0.189	0.855	0.386	0.468
17	0.866	0.431	0.436	-0.205	0.463	-0.667
18	0.633	0.414	0.219	-1.315	-1.559	0.244
19	1.069	0.827	0.242	-1.641	-1.327	-0.314
20	0.598	0.618	-0.020	-1.296	-1.043	-0.253
21	0.661	0.677	-0.016	-2.033	-1.285	-0.747
22	0.922	0.894	0.028	-0.723	-0.798	0.075
23	0.849	0.789	0.060	-0.961	-0.540	-0.421
24	0.789	0.785	0.004	-1.654	-0.903	-0.751
25	1.395	0.930	0.465	-0.031	-0.238	0.207
26	0.954	0.947	0.007	-0.865	-0.662	-0.203
27	0.632	0.974	-0.342	-1.190	-0.177	-1.013
28	0.759	0.428	0.330	-1.331	-2.098	0.767
29	0.745	0.842	-0.097	-1.023	-0.670	-0.353
30	0.886	1.076	-0.189	-0.750	-0.401	-0.349
31	1.262	0.896	0.366	-1.326	-1.490	0.164
32	1.207	0.781	0.426	-1.413	-1.533	0.121
33	1.773	0.967	0.806	-0.709	-0.519	-0.191
34	0.937	1.126	-0.189	-0.375	-0.297	-0.077
35	0.725	0.347	0.378	-1.277	-2.424	1.147
36	0.716	1.503	-0.787	-1.898	-1.432	-0.466
37	0.610	0.748	-0.139	-0.848	-0.630	-0.218
38	1.248	0.429	0.820	-0.127	-1.860	1.733
39	0.950	1.119	-0.169	-0.675	-0.234	-0.440

40	0.817	0.829	-0.012	-1.147	-1.382	0.235
41	0.758	0.816	-0.057	1.755	1.341	0.414
42	1.116	0.969	0.147	-0.564	-0.638	0.074
43	0.899	0.758	0.141	-2.203	-2.262	0.059
44	0.743	0.713	0.031	-0.172	0.118	-0.290
45	0.964	0.755	0.209	1.296	1.319	-0.023
46	1.000	1.577	-0.577	-0.582	-0.308	-0.274
47	1.423	0.801	0.622	-0.924	-1.029	0.105
48	0.607	0.960	-0.353	-0.847	-0.173	-0.674
49	0.554	0.753	-0.200	-1.623	-1.101	-0.522
50	0.836	0.930	-0.094	-2.109	-1.214	-0.895

Note. A negative difference indicates that the parameter is larger for the FIF students (FIE minus FIF = difference).

Table 8.

Item Parameters for the Social Studies Items – FIE versus FIF

Item	a parameter			b parameter		
	FIE	FIF	Difference	FIE	FIF	Difference
1	0.849	1.066	-0.217	-0.754	-1.012	0.258
2	0.483	0.670	-0.187	-1.845	-0.941	-0.904
3	0.559	0.529	0.030	0.258	0.708	-0.450
4	0.731	0.607	0.124	-1.231	-1.991	0.760
5	0.733	0.414	0.319	1.580	0.986	0.595
6	0.507	0.395	0.112	0.526	0.360	0.166
7	0.684	0.524	0.160	-1.140	-0.335	-0.805
8	1.636	1.104	0.532	0.720	0.192	0.528
9	0.901	0.927	-0.026	1.092	1.035	0.057
10	0.840	1.382	-0.542	-0.626	-0.780	0.154
11	0.925	0.764	0.161	0.311	-0.065	0.376
12	0.548	0.769	-0.222	0.091	-0.399	0.490
13	0.966	0.784	0.182	0.591	0.189	0.402
14	0.568	0.602	-0.034	-1.564	-2.072	0.508
15	1.707	1.012	0.696	0.526	0.722	-0.197
16	0.685	0.883	-0.198	-0.796	-0.443	-0.353
17	0.996	1.012	-0.016	-0.410	-0.592	0.182
18	0.862	1.035	-0.173	-1.234	-1.263	0.029
19	1.026	0.571	0.455	-0.326	-0.993	0.667
20	0.801	0.788	0.014	-1.303	-1.070	-0.233
21	0.788	0.925	-0.137	2.858	3.422	-0.564
22	0.615	0.633	-0.018	0.540	0.452	0.088
23	1.066	0.689	0.377	0.018	-0.369	0.387
24	1.011	0.466	0.545	-0.780	-0.737	-0.043
25	0.918	0.821	0.097	0.035	-0.807	0.842
26	0.982	0.962	0.020	-0.017	-0.253	0.236
27	0.650	5.203	-4.553	-0.828	0.068	-0.896
28	0.798	0.818	-0.021	-1.232	-0.759	-0.473
29	0.597	0.398	0.200	-0.805	-0.658	-0.147
30	1.006	1.287	-0.280	-0.102	-0.306	0.204
31	0.908	0.832	0.076	-0.276	-0.461	0.185
32	0.824	1.287	-0.462	0.252	0.447	-0.195
33	1.141	0.778	0.363	-0.246	-0.500	0.254
34	0.772	1.066	-0.294	-1.287	-0.717	-0.570
35	0.804	1.538	-0.734	-0.482	0.315	-0.797
36	1.000	1.129	-0.129	-1.115	-1.137	0.023
37	0.604	0.753	-0.149	-0.410	-0.573	0.163
38	0.614	0.526	0.087	-0.565	0.011	-0.576
39	0.722	0.791	-0.068	-0.116	0.557	-0.673

40	0.760	0.707	0.054	-0.881	-0.154	-0.727
41	1.395	0.685	0.710	0.864	0.910	-0.046
42	1.206	0.932	0.274	0.082	0.503	-0.421
43	0.493	0.542	-0.049	-0.606	-0.661	0.055
44	0.843	0.872	-0.029	-0.522	-0.445	-0.077
46	0.716	0.759	-0.044	0.318	0.165	0.154
47	1.307	0.696	0.611	0.925	0.707	0.219
48	0.963	0.992	-0.029	0.449	-0.212	0.661
49	0.717	0.948	-0.231	0.602	0.138	0.465
50	1.819	0.866	0.953	0.668	0.626	0.042

Note. A negative difference indicates that the parameter is larger for the Francophone students (French Immersion minus Francophone = difference).

Item 45 was removed from this table because BILOG was unable to estimate the b parameter for item 45.

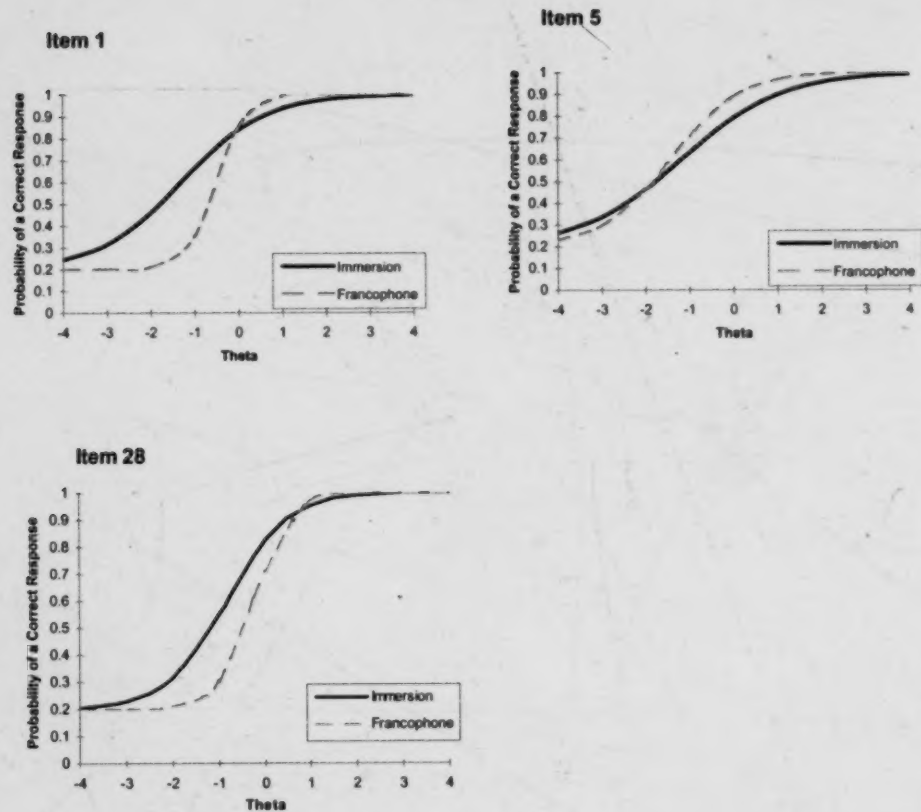


Figure 1.

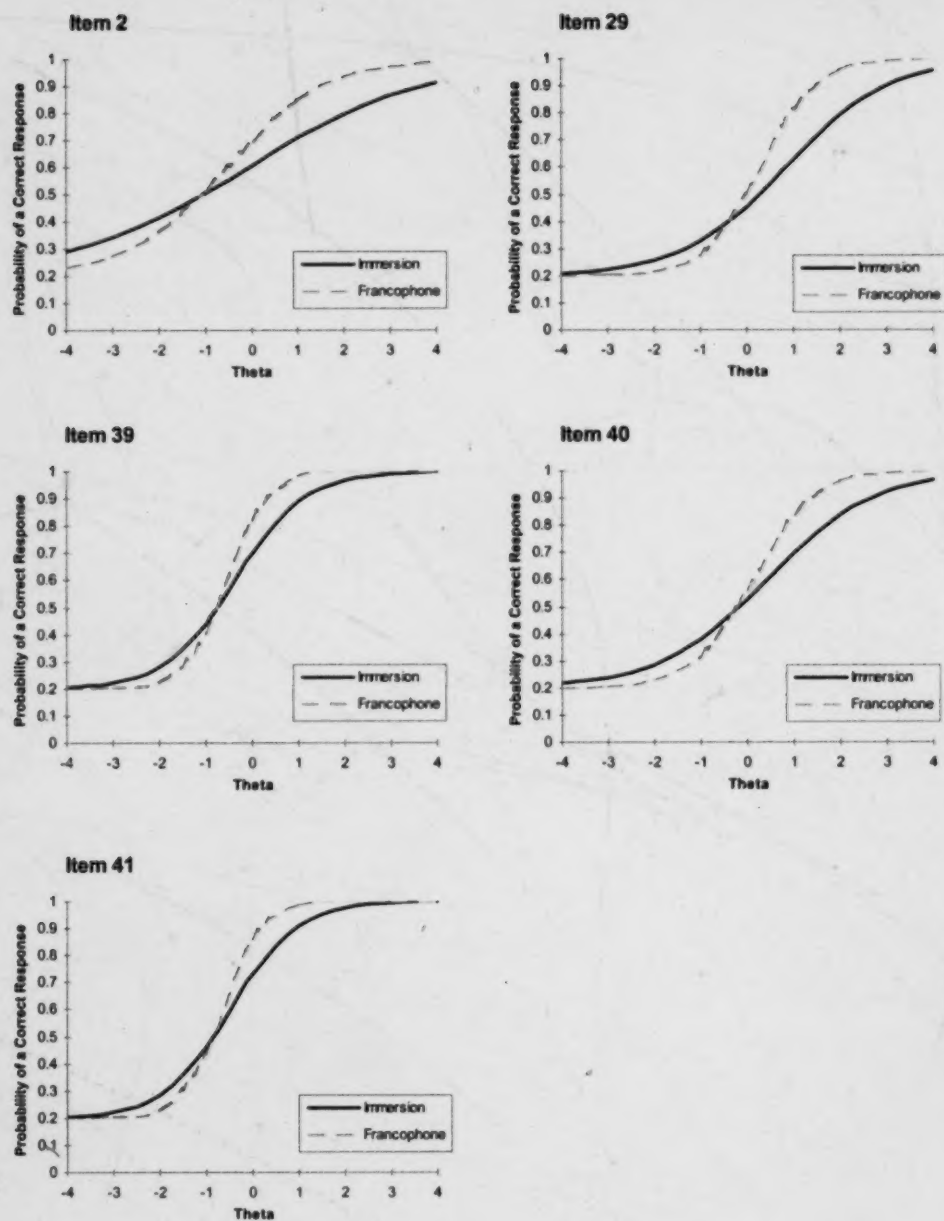


Figure 2.

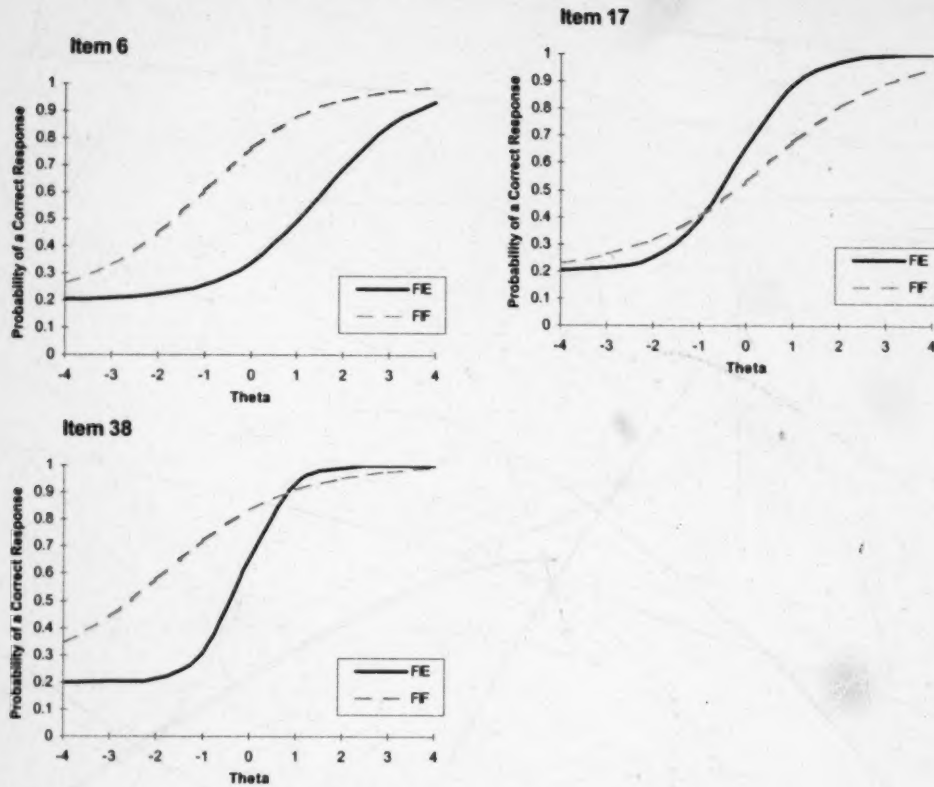
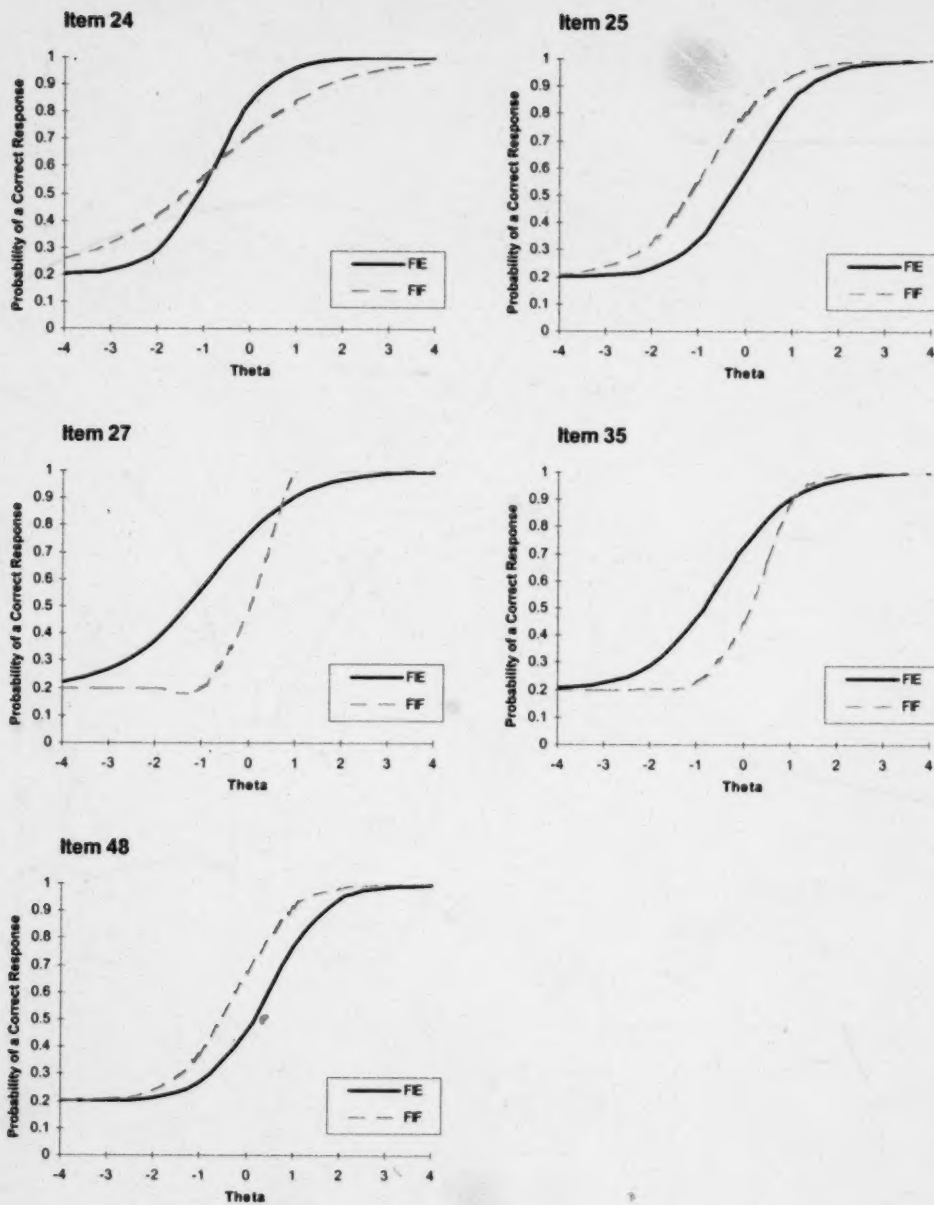


Figure 3.

**Figure 4.**